Dating the First Case of COVID-19 Epidemic from a Probabilistic Perspective

Zhouwang Yang[1], Yunhe Hu[1], Zhiwei Ding[1], Tiande Guo*[2]

[1]University of Science and Technology of China

[2]University of Chinese Academy of Sciences

* Corresponding author: tdguo@ucas.edu.cn (Tiande Guo)

ABSTRACT

In the early days of the epidemic of coronavirus disease 2019 (COVID-19), due to insufficient knowledge of the pandemic, inadequate nucleic acid tests, lack of timely data reporting, etc., the origin time of the onset of COVID-19 is difficult to determine. Therefore, source tracing is crucial for infectious disease prevention and control. The purpose of this paper is to infer the origin time of pandemic of COVID-19 based on a data and model hybrid driven method.

We model the testing positive rate to fit its actual trend, and use the least squares estimation to obtain the optimal model parameters. Further, the kernel density estimation is applied to infer the origin time of pandemic given the specific confidence probability.

By selecting 12 representative regions in the United States for analysis, the dates of the first infected case with 50% confidence probability are mostly between August and October 2019, which are earlier than the officially announced date of the first confirmed case in the United States on January 20, 2020. The experimental results indicate that the COVID-19 pandemic in the United States starts to spread around September 2019 with a high confidence probability.

In addition, the existing confirmed cases are also used in Wuhan City and Zhejiang Province in China to infer the origin time of COVID-19 and provide the confidence probability. The results show that the spread of COVID-19 pandemic in China is likely to begin in late December 2019.

KEYWORD

Pandemic; COVID-19; Testing positive rate; Infectious disease dynamic model; Least squares optimization; Kernel density estimation.

INTRODUCTION

In human history, it usually takes decades to explore the origin of infectious diseases and even until now people have not found all of the answers. Recently studying the origin, spread and evolution of coronavirus disease 2019 (COVID-19) in more than 200 countries and regions around the world has become a new research subject for global scientific community. Source tracking is crucial for infectious disease prevention and control. However, the process of scientific demonstration is complex for it requires a large amount of biological information and epidemiological evidence to converge into a mutually supportive evidence chain, which is time-consuming and uncertain. A series of previous studies showed that the United States, Spain, France,

Italy, Brazil and other countries had been attacked by the coronavirus before its outbreak in China.

The main task of disease tracing is to find the first case. Though the first known case in Wuhan, China is the first confirmed case reported, it does not mean it is the first case people seek for. From the human history of fighting against infectious diseases, it is difficult to find a successful precedent for tracing the origin yet.

The primary method for origin tracing of COVID-19 is molecular traceability [1]. First of all, a global coronavirus information database is needed to further integrate genomic, epidemiological and clinical data. Secondly, based on the integration and analysis of molecular data and epidemiological data, it is able to systematically study the correlation and law between this series of coronavirus and various exposed factors, which provide an important reference for traceability.

Epidemic spread is a complex process involving many factors [2], some of which are difficult to figure out. However, the epidemic data imply the comprehensive influence of these factors. Theoretically, by analyzing these big data, the law of epidemic spread can also be obtained. Therefore, another method for origin tracing of COVID-19 is based on big data analysis. Combined with mathematical model and artificial intelligence technology, qualitative and quantitative analysis of infectious diseases can reveal the epidemic law of infectious diseases and detect the origin and development trend of infectious diseases. There are many studies on predicting forward using epidemic model and data at home and abroad [3-7], but there are few studies on tracing backward by establishing mathematical models and using big data analysis methods [8,9].

In the early days of COVID-19, most countries including the United States and China lacked basic knowledge of the epidemic situation and nucleic acid detection was not in place. There were other problems such as lack, lag or distortion in the data released, which made it even more difficult to determine the origin time of the epidemic onset. To this end, we collect daily epidemic data of the U.S. including the number of newly confirmed people, the number of new deaths, the number of Nucleic Acid Amplification Tests (NAATs) and the positive rate of tests. We analyze the characteristics of various data and finally choose the number of nucleic acid tests and the testing positive rate of each state in the U.S. as the modeling data. According to the classical infectious disease model and statistical methods, an optimization model is established. The model parameters are obtained by using least squares estimation. The origin time of epidemic situation in selected states of the U.S. is inferred, and the time with corresponding probability 0.5, 0.6, 0.7 and 0.8 of the first infection, 50 infections and 100 infections in these states are obtained by kernel density estimation.

RESULTS

Data description

The main data used for modeling in this study are the daily testing positive rate, that is, the daily proportion of the number of positive nucleic acid tests accounts for the total number of nucleic acid tests of COVID-19. The data on the total number of tests and the number of positive tests of each state in the U.S. come from the official website of the United States Department of Health and Human Services [10]. By observing the early testing positive rate curves of more than 50 states in the U.S., it is found that 13 states and District of Columbia (mainly in the Northeast) where the excess mortality reached the peak earlier in 2020 [11] share the same pattern of change, that is, the testing positive rate rises to the peak rapidly after a short fluctuation.

Table 1 shows the cumulative number of tests, the population and the percentage of total tests in

the population of the 13 states and District of Columbia in the U.S. at the peak of the testing positive rate. When the testing positive rate in New Jersey and Vermont reached the peak, the cumulative number of tests was less than 1,000 and the test ratio was too small, so these two states are not considered in the following analysis.

Table 1. Cumulative Number of Tests, Population and Test Ratio of the 13 States and District of Columbia in the U.S.

Region	Cumulative Number of Tests	Population	Test Ratio
New Jersey	158	8882190	0.000018
Vermont	715	623989	0.001146
Virginia	18194	8535519	0.002132
Michigan	50905	9986857	0.005097
New Hampshire	12001	1359711	0.008826
Louisiana	45567	4648794	0.009802
Connecticut	44479	3565287	0.012476
New York	247165	19453561	0.012705
Pennsylvania	168746	12801989	0.013181
Maryland	86939	6045680	0.01438
District of Columbia	12256	705749	0.017366
Massachusetts	126162	6892503	0.018304
Delaware	22636	973764	0.023246
Rhode Island	44018	1059361	0.041551

Take Maryland as an example (Figure 1). The black dots represent the calculated daily testing positive rate, and the red line represents the values after 15-day smoothing (7 days each before and after that very day). The purpose of smoothing is to reduce the impact of data fluctuations. The following testing positive rate refers to the smoothed testing positive rate unless specified otherwise. The positive rate in Maryland began to increase from 8% on March 18, and reached a peak of nearly 30% on April 15, after which the positive rate started to decline.

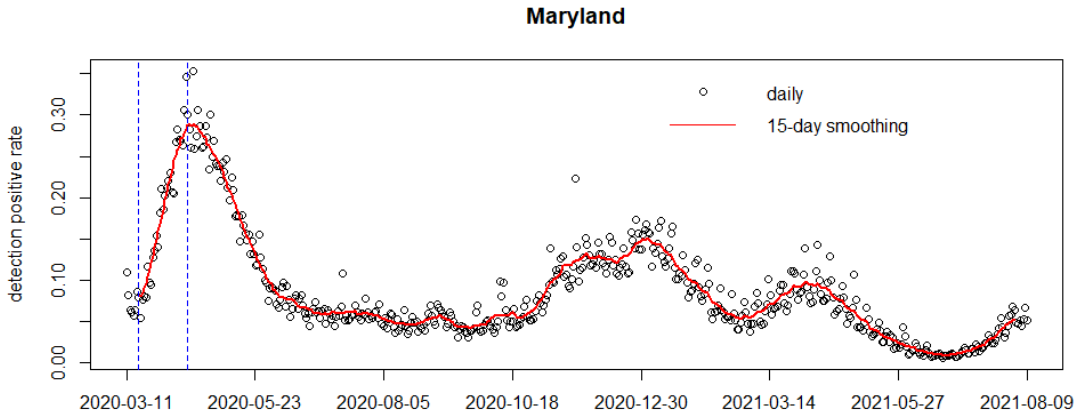
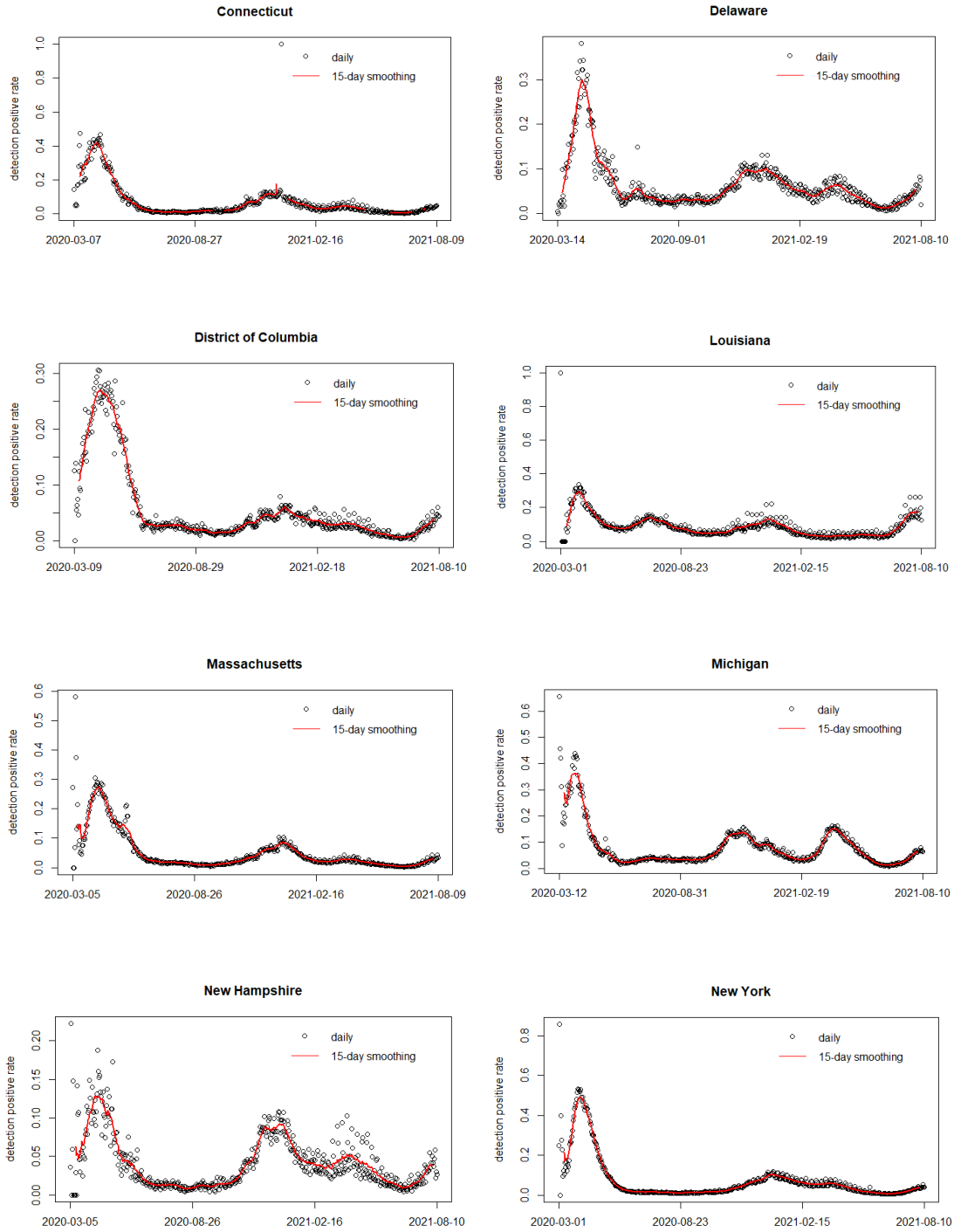


Figure 1. Testing Positive Rate of Maryland

The positive rates of the remaining 10 states and District of Columbia have the same characteristics as that of Maryland, as shown in Figure 2. All states began to open commercial nucleic acid tests around March 15 and before that, due to the limit of detection level and inadequate number of tests, the positive rate may fluctuate to some extent and cannot represent the actual situation. Therefore, only the steadily increasing sequence from the first valley to the first peak is selected for modeling.



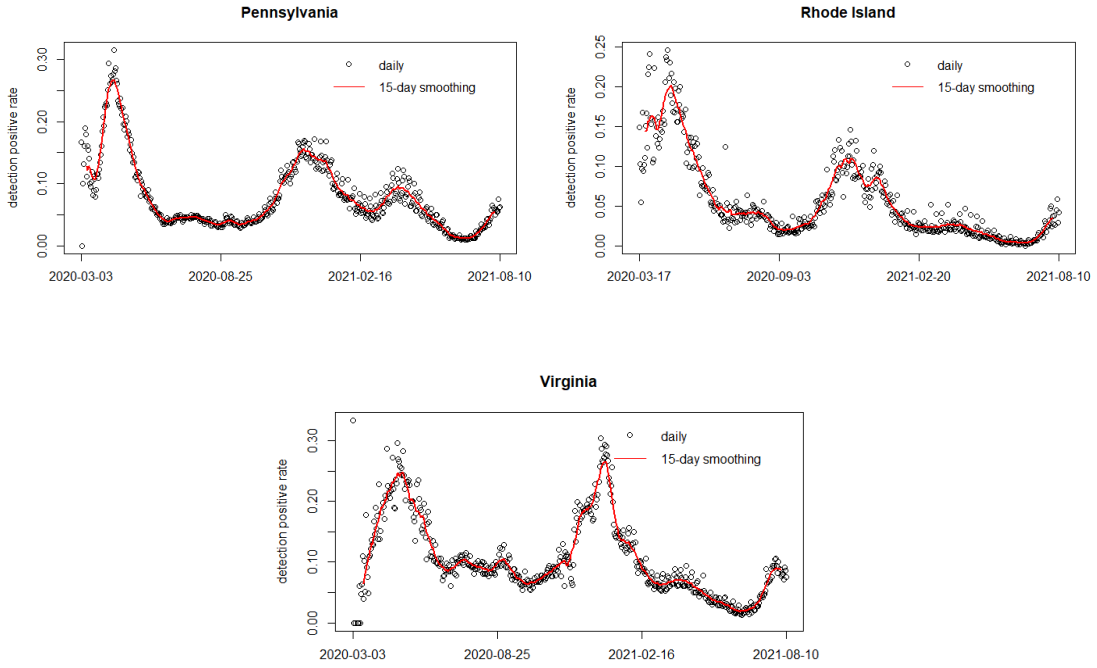


Figure 2. Testing Positive Rates of 10 states and District of Columbia

The above-mentioned 11 states and District of Columbia share the same characteristic that the testing positive rate rises to a peak not long after the beginning, due to the approximately natural propagation in the early stage of the epidemic in the U.S. If for each state, the infectious disease model we build can fit the first rising part of the positive rate well, it indicates that the model can accurately reflect the spread of the epidemic within time, then we can trace the origin of the epidemic by looking backwards in history.

Date tracing process

The process of tracing the origin time of the pandemic is mainly divided into three steps.

Step 1. Perform a 15-day smoothing process on the daily positive rate of the target district to reduce the impact of random noise. If there are abnormal fluctuations in the previous period, drop this part of the data, and only choose the sequence corresponding to the first stably rising period. The time interval is denoted as T, whose length is denoted as τ, and the endpoints on both sides correspond to the trough time and the crest time respectively.

Step 2. Take the 14 consecutive days of interval T as the fitting data $y_i, i = 1, \dots, 14$, use the two-parameter exponential epidemic model of the testing positive rate to get the fitting function $\hat{y}$, and record the fitting accuracy index MAPE (Mean Absolute Percentage Error)

$$\frac{1}{14}\sum_{i=1}^{14}\frac{|\hat{y}_i - y_i|}{y_i}.$$

Denote the number of the people engaged in NAATs in the target district as M. Extend the positive rate fitting function $\hat{y}$ towards history, and solve for the time $\bar{t}^1$ when $\hat{y}(t)\mathrm{M} = 1$, that is, the occurring time of first case of the target district. Similarly, solve for the time $\bar{t}^{50}$ when $\hat{y}(t)\mathrm{M} = 50$ and the time $\bar{t}^{100}$ when $\hat{y}(t)\mathrm{M} = 100$, which represent the occurring time of 50

cases and 100 cases in the target district respectively. Since $\hat{y}(t) \to 0, t \to -\infty$, the above equation must have a solution.

Step 3. Take 14 days as the size of the fitting window and 1 as the step size, and perform sliding sampling on the interval T. Repeat step 2 for each window to obtain τ-13 retrospective dates and MAPE values. Apply the kernel density estimation to obtain the probability distribution of the origin time, and calculate the average MAPE as the evaluation index of the overall fitting accuracy.

Origin time of 11 states and District of Columbia in the U.S.

Take Maryland as an example to trace the origin of the epidemic. The rising period of the testing positive rate in Maryland is from March 18, 2020 to April 15, 2020. The observation data, transmission model data, and retrospective data are shown in Figure 3. The time interval between the blue dotted lines corresponds to March 18, 2020 to April 15, 2020, and the time interval between the gray dotted lines is one of the sliding data windows used for fitting.

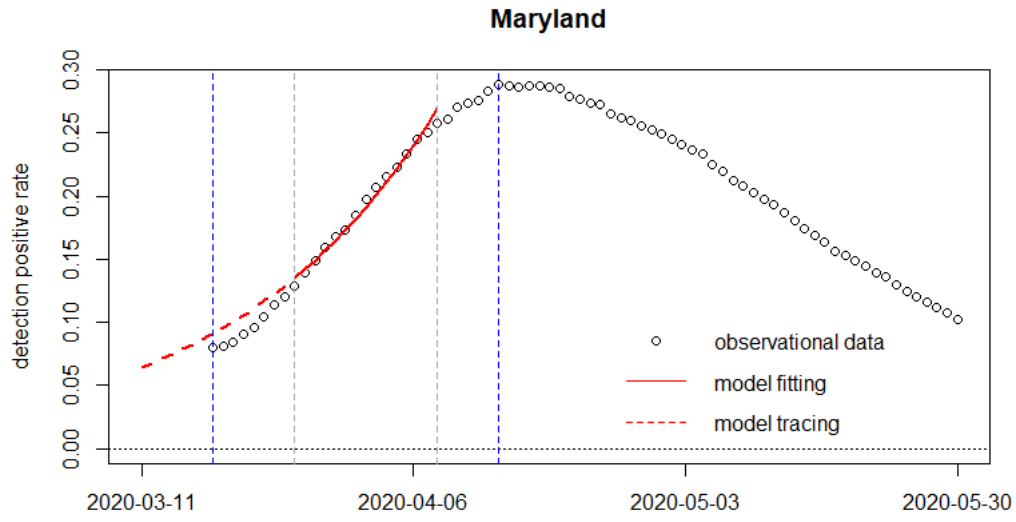
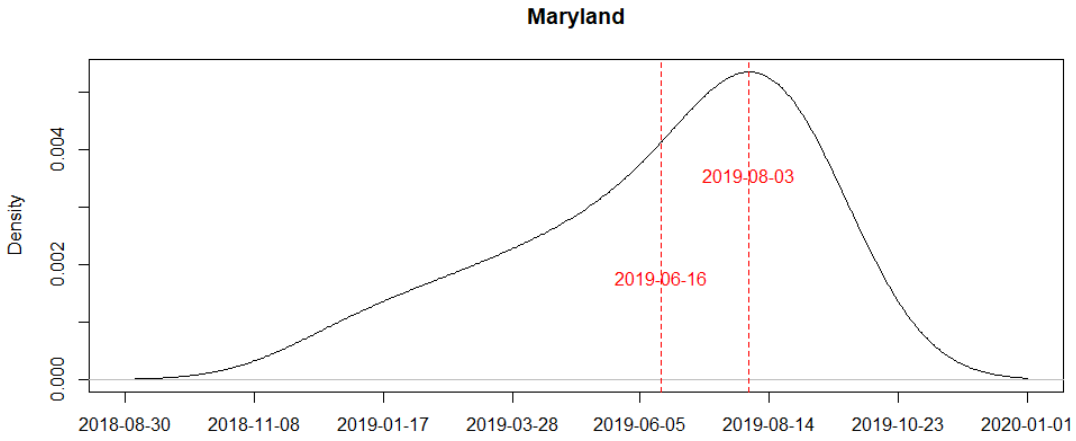
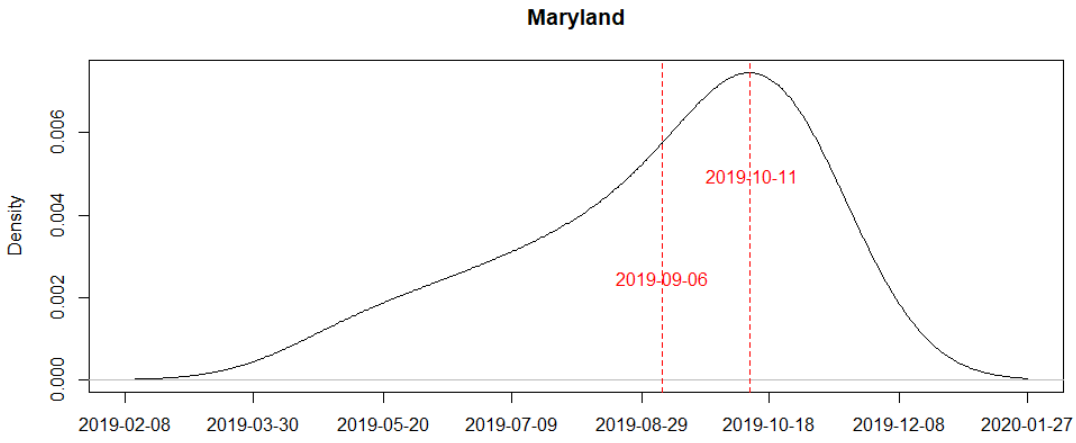


Figure 3. Modeling the testing positive rate in Maryland and tracing the Origin date

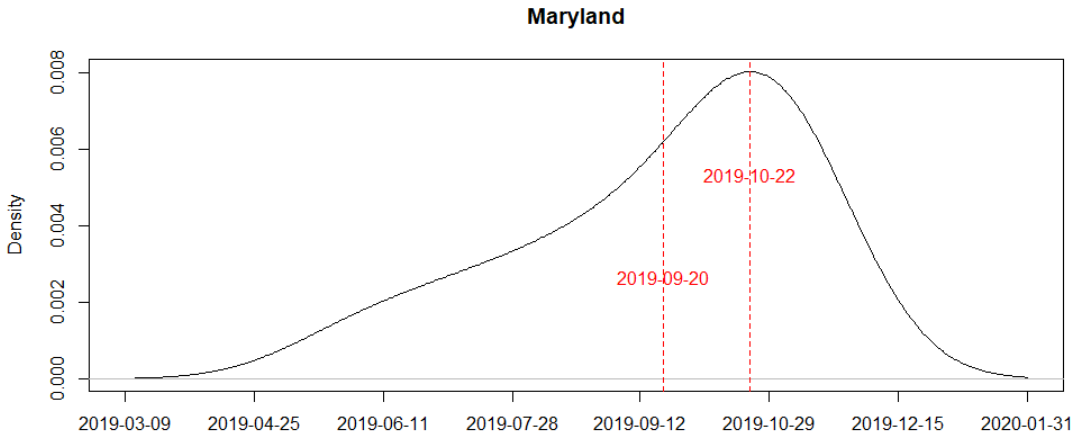
Through the sliding of the fitting window, several inferred dates of the first case in Maryland are obtained and the corresponding probability density is shown in the Figure 4A. In the same way, dates of 50, 100 cases are inferred and their corresponding probability densities are shown in the Figure 4B and 4C. The two red lines in figures represent the mean line (left) and the density peak line (right).



(a) first case



(b) 50 cases



(c) 100 cases

Figure 4. Tracing dates of the first case, 50 cases, 100 cases and corresponding probability density for Maryland

Table 2 shows the results of inferring origins of the COVID-19 pandemic for 11 states and District of Columbia in the U.S., including dates of the first infection, 50 infections, and 100 infections given the probability of 50%, 60%, 70%, and 80%, respectively. Taking Maryland as an example, the probabilities that the first infection occurred before 2019-09-22, 2019-10-06, 2019-10-19, and 2019-11-02 are 50%, 60%, 70%, and 80%, respectively.

Table 2. Dates and corresponding probabilities of the first, 50, and 100 outbreaks in 12 regions of the United States

Region	First Case				50 Cases				100 Cases				Average MAPE
	50%	60%	70%	80%	50%	60%	70%	80%	50%	60%	70%	80%	
Rhode Island	2019/4/26	2019/5/7	2019/5/18	2019/5/29	2019/9/28	2019/10/5	2019/10/10	2019/10/17	2019/10/26	2019/10/31	2019/11/6	2019/11/11	0.0116
Connecticut	2019/5/12	2019/5/25	2019/6/6	2019/6/20	2019/9/20	2019/9/27	2019/10/5	2019/10/13	2019/10/13	2019/10/20	2019/10/27	2019/11/2	0.013
Michigan	2019/8/14	2019/8/27	2019/9/8	2019/9/21	2019/11/14	2019/11/21	2019/11/28	2019/12/6	2019/11/30	2019/12/6	2019/12/13	2019/12/20	0.0159
District of Columbia	2019/8/29	2019/9/21	2019/10/9	2019/10/27	2019/12/14	2019/12/25	2020/1/3	2020/1/12	2020/1/1	2020/1/11	2020/1/18	2020/1/25	0.0228
Maryland	2019/9/22	2019/10/6	2019/10/19	2019/11/2	2019/12/9	2019/12/17	2019/12/25	2020/1/2	2019/12/23	2019/12/30	2020/1/6	2020/1/13	0.0211
Pennsylvania	2019/9/22	2019/9/26	2019/9/29	2019/10/2	2019/12/3	2019/12/6	2019/12/7	2019/12/9	2019/12/16	2019/12/18	2019/12/20	2019/12/22	0.0209
Massachusetts	2019/10/5	2019/10/18	2019/10/30	2019/11/12	2019/12/13	2019/12/21	2019/12/28	2020/1/5	2019/12/24	2019/12/31	2020/1/7	2020/1/14	0.0422
New York	2019/10/12	2019/10/16	2019/10/20	2019/10/23	2019/12/7	2019/12/10	2019/12/12	2019/12/15	2019/12/17	2019/12/19	2019/12/21	2019/12/23	0.0378
New Hampshire	2019/10/22	2019/10/29	2019/11/4	2019/11/11	2020/1/20	2020/1/23	2020/1/26	2020/1/29	2020/2/4	2020/2/7	2020/2/9	2020/2/11	0.0277
Virginia	2019/10/23	2019/11/3	2019/11/15	2019/11/27	2020/1/9	2020/1/14	2020/1/20	2020/1/26	2020/1/23	2020/1/27	2020/2/1	2020/2/5	0.0168
Louisiana	2019/11/4	2019/11/15	2019/11/26	2019/12/7	2020/1/1	2020/1/7	2020/1/13	2020/1/20	2020/1/11	2020/1/16	2020/1/22	2020/1/28	0.0777
Delaware	2019/11/30	2019/12/9	2019/12/18	2019/12/27	2020/1/29	2020/2/2	2020/2/7	2020/2/12	2020/2/8	2020/2/12	2020/2/16	2020/2/20	0.0306

The average MAPE for modelling each state's testing positive rate is less than 5%, indicating that the models are of high accuracy. In addition, the rising period of the positive rate is short in several states, thus the length of the fitting window is appropriately reduced to obtain more results of inferring origins, ensuring the accuracy of the kernel density estimation.

The above uses the cumulative number of positive tests to conduct into the origins of COVID-19. Since the actual number of positive people is much larger than the number of positive tests due to the limitation of testing, the former inferences are relatively conservative inference, that is, the inferred dates of origins are relatively late. The authoritative study has shown that the actual infected cases of COVID-19 in the United States are between 3 and 20 times the number of confirmed cases [12], which means that the early detection of the epidemic in the United States is obviously insufficient, resulting in that the number of infected cases is seriously underestimated. To this end, we have expanded the cumulative number of people involved in the nucleic acid test up to the peak of the positive rate to 3, 5, 10, 15, and 20 times, respectively, and then we can infer the earlier date of the onset of the epidemic. The testing rate in Maryland at the peak of the positive is at a medium level among the selected regions. As a typical representative example, we will expand Maryland's cumulative number of tests before the peak of positive rate rises to 3, 5, 10, 15, and 20 times, respectively, the corresponding dates of origin and probabilities are shown in Table 3.

Table 3. The date of origin of the epidemic and its corresponding probabilities for the cumulative number of people detected in Maryland expanded by different multiples

Multiple	First Case				50 Cases				100 Cases			
	50%	60%	70%	80%	50%	60%	70%	80%	50%	60%	70%	80%
3	2019/9/1	2019/9/16	2019/9/30	2019/10/16	2019/11/17	2019/11/27	2019/12/6	2019/12/16	2019/12/1	2019/12/10	2019/12/18	2019/12/27
5	2019/8/21	2019/9/7	2019/9/22	2019/10/8	2019/11/7	2019/11/18	2019/11/27	2019/12/8	2019/11/21	2019/11/30	2019/12/10	2019/12/19
10	2019/8/8	2019/8/25	2019/9/10	2019/9/27	2019/10/24	2019/11/5	2019/11/16	2019/11/27	2019/11/7	2019/11/18	2019/11/27	2019/12/8
15	2019/7/31	2019/8/17	2019/9/3	2019/9/20	2019/10/16	2019/10/29	2019/11/9	2019/11/20	2019/10/30	2019/11/10	2019/11/21	2019/12/2
20	2019/7/25	2019/8/13	2019/8/29	2019/9/17	2019/10/10	2019/10/23	2019/11/4	2019/11/16	2019/10/24	2019/11/5	2019/11/16	2019/11/27

Origin time of Wuhan City and Zhejiang Province in China

The number of 'existing confirmed cases' is defined as the 'cumulative number of confirmed cases' minus the 'sum of cumulative number of recovered cases and cumulative number of deaths'.

As China adopts a mass testing strategy of epidemic prevention and control [13], the testing positive rate maintains a very low level in most time, which is not suitable for modeling. However, because of this strategy, the number of existing confirmed cases in China is closer to the actual number of infections. Therefore, we directly use the number of existing confirmed cases to replace the testing positive rate, and select Wuhan City and Zhejiang Province, China, as the two representative regions to trace the origin of the COVID-19.

The changes in the existing confirmed cases in Wuhan City and Zhejiang Province [14,15] are shown in Figure 5. As for Wuhan (Figure 5A), we find that the number of existing confirmed cases increased sharply on February 12, 2020. This is mainly caused by the revision of the definition of confirmed cases. Specifically, the number of clinically diagnosed cases is also included into the number of confirmed cases. The number of existing confirmed cases peaks on February 18, 2020. With respect to Zhejiang (Figure 5B), the number of existing confirmed cases peaks on February 7, 2020.

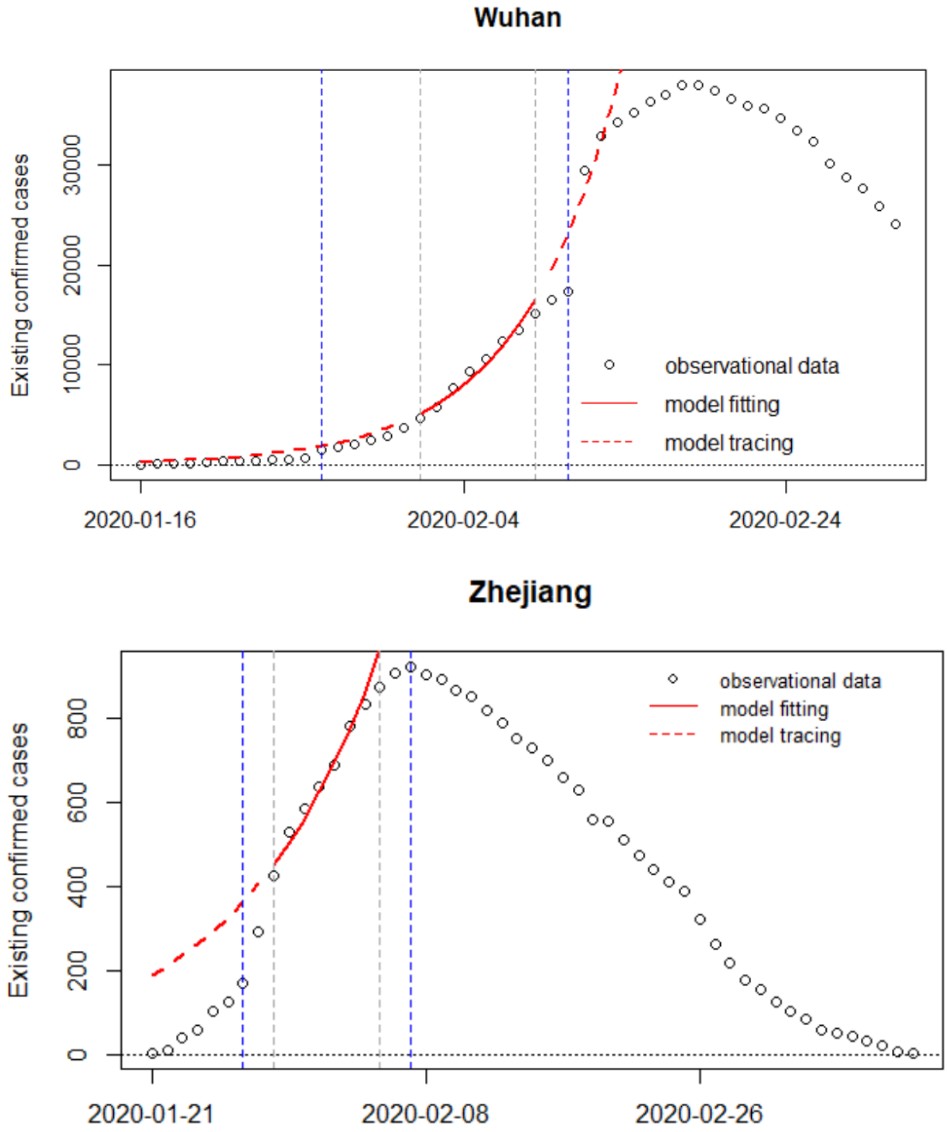
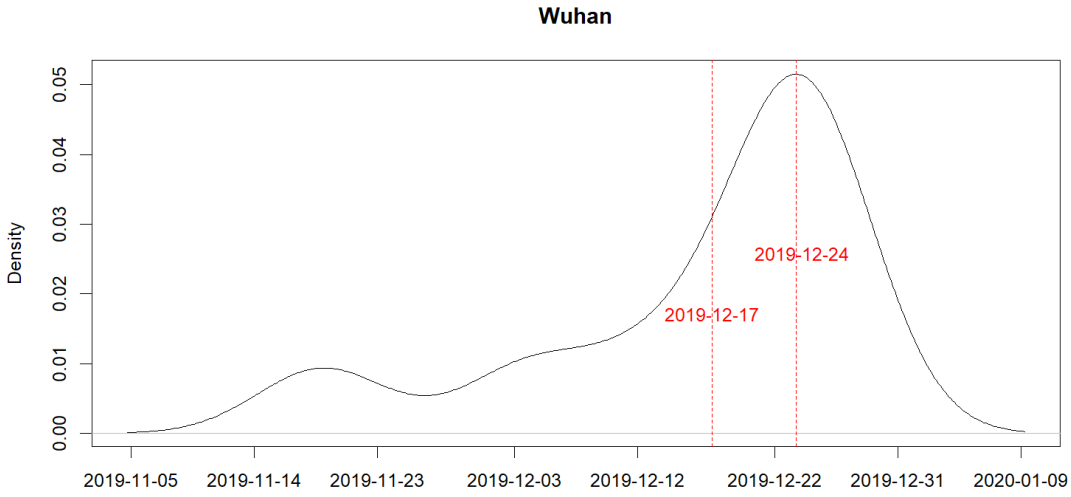


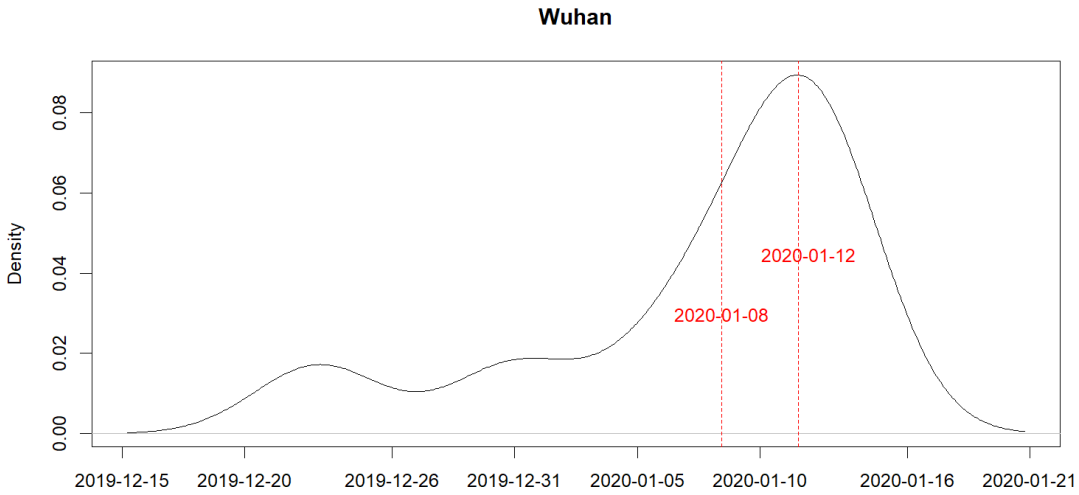
Figure 5. Modeling the number of existing confirmed cases and tracing the Origin time in Wuhan and Zhejiang, China

In order to improve the reliability of results, we change the time interval and the size of sliding window to conduct multiple numerical experiments, and select the model with the smallest MAPE as the final model. For Wuhan, we take January 27, 2020 to February 11, 2020, and January 27, 2020 to February 18, 2020 as the fitting time intervals. Meanwhile, we change the window size. With regard to Zhejiang, we take January 22, 2020 to February 7, 2020, January 23, 2020 to February 7, 2020, January 24, 2020 to February 7, 2020, and January 25, 2020 to February 7, 2020 as the fitting time interval, respectively. The model results of two regions are listed in Table 5. The model selected for furtherly inferring the Origin time is marked with the star in the MAPE column (Table 5).

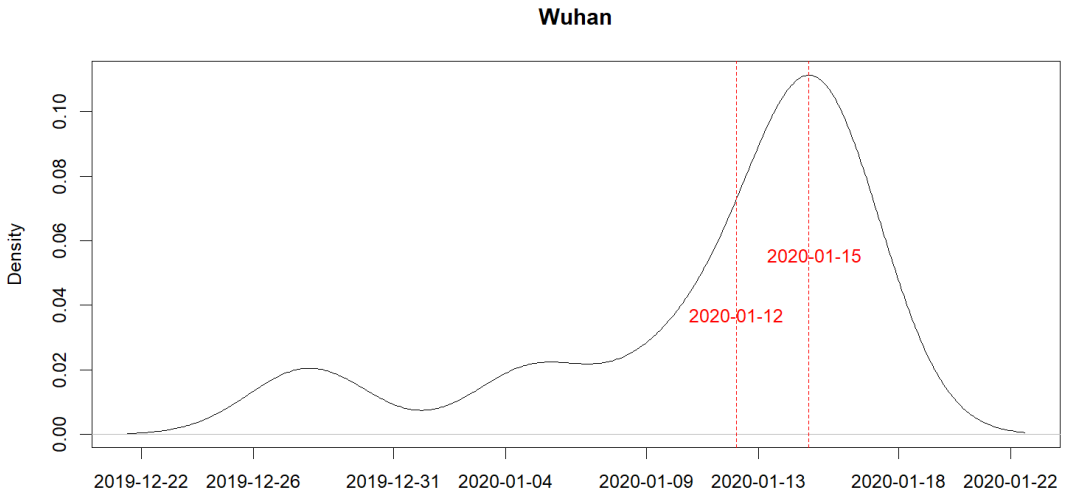
Through the sliding of the fitting window, multiple inferred dates of the first case, 50 cases, 100 cases in Wuhan are obtained and the corresponding probability density is shown in Figure 6.



(a) first case



(b) 50 cases



(c) 100 cases

Figure 6. Tracing dates of the first case, 50 cases, 100 cases and corresponding probability density for Wuhan

Table 5 shows that the probabilities that the first infection occurred in Wuhan, China before December 20, 2019, December 22, 2019, December 24, 2019, and December 26, 2019 are 50%, 60%, 70%, and 80%, respectively. The probabilities that the first infection occurred in Zhejiang, China before December 23, 2019, December 31, 2019, January 6, 2020, and January 14, 2020 are 50%, 60%, 70%, and 80%, respectively.

Table 5. Dates and corresponding probabilities of the first case, 50 cases and 100 cases in Wuhan and Zhejiang, China

Region	First Case				50 Cases				100 Cases				Average MAPE	Start	End	Window Size
	50%	60%	70%	80%	50%	60%	70%	80%	50%	60%	70%	80%				
Wuhan	2019/12/12	2019/12/14	2019/12/16	2019/12/18	2020/1/5	2020/1/6	2020/1/7	2020/1/9	2020/1/9	2020/1/10	2020/1/11	2020/1/12	0.0964	2020/1/27	2020/2/18	14
Wuhan	2019/12/11	2019/12/15	2019/12/19	2019/12/23	2020/1/4	2020/1/7	2020/1/9	2020/1/11	2020/1/8	2020/1/11	2020/1/12	2020/1/14	0.0690	2020/1/27	2020/2/18	7
Wuhan	2019/12/20	2019/12/22	2019/12/24	2019/12/26	2020/1/9	2020/1/10	2020/1/11	2020/1/13	2020/1/13	2020/1/14	2020/1/15	2020/1/16	0.04+	2020/1/27	2020/2/11	7
Zhejiang	2019/12/26	2020/1/2	2020/1/9	2020/1/16	2020/1/17	2020/1/20	2020/1/22	2020/1/24	2020/1/20	2020/1/22	2020/1/23	2020/1/25	0.1040	2020/1/22	2020/2/7	7
Zhejiang	2019/12/23	2019/12/31	2020/1/6	2020/1/14	2020/1/16	2020/1/19	2020/1/21	2020/1/24	2020/1/20	2020/1/22	2020/1/23	2020/1/25	0.0874+	2020/1/23	2020/2/7	7
Zhejiang	2019/12/18	2019/12/26	2020/1/3	2020/1/11	2020/1/14	2020/1/18	2020/1/21	2020/1/23	2020/1/19	2020/1/21	2020/1/23	2020/1/25	0.0900	2020/1/24	2020/2/7	7
Zhejiang	2019/12/13	2019/12/21	2019/12/30	2020/1/7	2020/1/12	2020/1/16	2020/1/19	2020/1/22	2020/1/18	2020/1/20	2020/1/22	2020/1/25	0.0881	2020/1/25	2020/2/7	7

Conclusion

Based on the infectious disease transmission model and big data analysis method, this paper establishes an optimization model. Using the daily data released by the 12 representative regions of the United States, the model parameters are obtained separately, and then dates of first case, 50 cases and 100 cases of COVID-19 infection are inferred with corresponding probabilities. For the 12 representative regions, the dates of the first infection with probability 50% are mostly between August and October 2019, the earliest is April 26, 2019 for Rhode Island, and the latest is November 2019 for Delaware, which are all earlier than January 20, 2020, the officially announced date of the first confirmed case in the United States. The calculation results show that the COVID-19 epidemic in the United States has a high probability of beginning to spread around September 2019.

According to this model, we use the daily existing number of confirmed diagnoses in Wuhan City, China and Zhejiang Province, China to obtain the model parameters and infer the infection time of

first case, 50 cases and 100 cases with corresponding probabilities. For Wuhan, the date of the first case of COVID-19 with probability 50% is inferred as December 20, 2019, and the date of the first case in Zhejiang is inferred as December 23, 2019. The calculation results show that the COVID-19 epidemic in China has a high probability of beginning to spread in late December 2019.

If the detection data in the early stage of epidemic of other countries or regions are relatively accurate, this method can be used to infer the Origin time of the epidemic and provide the date of the first case or certain cases under a given probability.

METHODS

Epidemic model

The classic infectious disease dynamic model assumes that the number of infected persons will increase exponentially in the early stage of epidemic under the condition of non-intervention and approximately natural transmission, and thus intuitively presents a J curve. This assumption is consistent with the early epidemic situation of the U.S. Literature [16] proposed the following infectious disease transmission model

$$N(t) = N(t_0)\exp\{a_t(t-t_0)\}, \tag{1}$$

where $N(t)$ is the number of existing infections at time t, $N(t_0)$ and t_0 are constants, and a_t varies with time t. Due to the short initial spread of the epidemic, the virus has not mutated yet, we can assume that a_t does not change over time and is denoted as a. After simplifying the model (1), a two-parameter exponential model of the number of existing infections is obtained

$$N(t) = e^{b+at}. \tag{2}$$

Denote S as the set of cumulative test population up to the first peak of testing positive rate in the target district. Then the population participating in the test every day can be regarded as a random sampling of the set S, and the test positive rate represents the infection rate of the set S. The positive rate is multiplied by M (the number of elements in S) to get the number of existing infections in the test population, which is much lower than the number of existing infections in the target district during the same period. The curve of the testing positive rate synchronizes with the curve of the existing number of infected persons, and the only difference between them is a constant multiple, so the model (2) is also applicable to the testing positive rate. Denote the testing positive rate in the target district as y. From the above two-parameter exponential model (2), the following model can be obtained

$$y(t) = \mathrm{e}^{c_0+c_1t}, \tag{3}$$

where c_0, c_1 are model parameters to be determined, which can be estimated from the observational data of testing positive rate.

Least squares optimization

Assuming that a total of n days of observational data $(t_i, y_i), i = 1, \dots, n$ are collected, in order to estimate the parameters c_0 and c_1, the following least squares optimization model is established

$$\min_{c_0,c_1} \sum_{i=1}^{n} (y_i - e^{c_0+c_1t_i})^2. \tag{4}$$

To simplify the calculation, first take the natural logarithm of both sides of the model (3) to obtain $\log y = c_0 + c_1t$, and then for the transformed data $(t_i, \log y_i), i = 1, \dots, n$, we reformulate the least squares optimization model

$$\min_{c_0,c_1} \sum_{i=1}^{n} (\log y_i - c_0 - c_1t_i)^2. \tag{5}$$

Solve model (5) to get the optimal solution $(\hat{c}_0, \hat{c}_1)$, thus obtaining the epidemic spread model $\hat{y}(t) = \mathrm{e}^{\hat{c}_0 + \hat{c}_1 t}$.

Kernel density estimation

Kernel density estimation, as a non-parametric estimation method, can be applied to estimate the unknown probability distribution without prior knowledge. The principle is that if a certain number appears in observation, it can be considered that the probability density of this number is relatively large, the probability density of the number close to it will also be relatively large, and the probability density of the number far away from it will be relatively small. Therefore, a function that satisfies the above conditions can be used to approximate the probability density for each observed number, and then sum all the functions to obtain the probability density function after normalization.

Assuming that several possible Origin times of the target district are calculated as $x_i, i = 1, \ldots, m$ according to different data fitting intervals, the probability density of the Origin date at x is

$$\hat{f}_h(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(x - \frac{x_i}{h}\right),$$

where the kernel function K satisfies $\int K(x)\mathrm{d}x = 1$, and the smoothing parameter h is called bandwidth. The kernel function is generally a symmetric and unimodal probability density function. Here, the commonly used Gaussian kernel is selected as

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

and the choice of bandwidth follows Silverman's rule of thumb [17].

FUNDING

The work is supported by Anhui Center for Applied Mathematics, the NSF of China (No. 11871447).

REFERENCES

[1] Shan K J, Wei C, Wang Y, Huan Q, Qian W. Host-specific asymmetric accumulation of mutation types reveals that the origin of SARS-CoV-2 is consistent with a natural process[J]. The Innovation, 2021, doi: https://doi.org/10.1016/j.xinn.2021.100159.

[2] Wu R, Ai S, J Cai, et al. Predictive Model and Risk Factors for Case Fatality of COVID-19: A Cohort of 21,392 Cases in Hubei, China[J]. The Innovation, 2020, 1(2): 100022.

[3] Shen, M., Peng, Z., Xiao, Y. and Zhang, L. Modeling the Epidemic Trend of the 2019 Novel Coronavirus Outbreak in China. The Innovation, 2020, 1(3): 100048.

[4] Sun H, Qiu Y, Yan H, Huang Y, Zhu Y and Chen S. Tracking and Predicting COVID-19 Epidemic in China Mainland. medRxiv, 2020.

[5] Chen Y, Lu P, Chang C and Liu T. A Time-Dependent SIR Model for COVID-19 With Undetectable Infected Persons. IEEE Transactions on Network Science and Engineering, 2020, 7(4): 3279-3294.

[6] Giordano G, Blanchini F and Bruno R. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nature Medicine, 2020.

[7] Zhang Y, You C, Cai Z, et al. Prediction of the COVID-19 outbreak in China based on a new stochastic dynamic model. Scientific Reports, 2020.

[8] Roberts DL, Rossman JS, Jarić I. Dating first cases of COVID-19. PLoS Pathog, 2021, 17(6): e1009620.
[9] Zhang C, Wang M. Origin time and epidemic dynamics of the 2019 novel coronavirus. bioRxiv, 2020.
[10] https://healthdata.gov/dataset/COVID-19-Diagnostic-Laboratory-Testing-PCR-Testing/j8mb-icvb
[11] https://www.nytimes.com/interactive/2020/us/covid-death-toll-us.html
[12] Wu S L, Mertens A N, Crider Y S, et al. Substantial underestimation of SARS-CoV-2 infection in the United States[J]. Nature Communications, 2020, 11(1): 4507.
[13] Shen M, Xiao Y, Zhuang G, Li Y and Zhang L. Mass testing—An underexplored strategy for COVID-19 control. The Innovation, 2021.
[14] https://github.com/CSSEGISandData/COVID-19
[15] http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml
[16] Huang N E, Qiao F. A data driven time-dependent transmission rate for tracking an epidemic: a case study of 2019-nCoV[J]. Science Bulletin, 2020, 65(6): 425-427.
[17] Silverman B W. Density Estimation for Statistics and Data Analysis[M]. Chapman and Hall, 1986.